CLUSTERING OF CONTEXT DEPENDENT SPEECH UNITS FOR MULTILINGUAL SPEECH RECOGNITION

Bojan Imperl
University of Maribor
Smetanova 17, 2000 Maribor, SLOVENIA
E-mail: bojan.imperl@uni-mb.si

ABSTRACT

The paper addresses the problem of designing a language independent phonetic inventory for the speech recognisers with multilingual vocabulary. A new clustering algorithm for the definition of multilingual set of triphones is proposed. The clustering algorithm bases on a definition of a distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. The monophone similarity estimation method based on the algorithm of Houtgast. The clustering algorithm is integrated in a multilingual speech recognition system based on HTK V2.1.1. The experiments were based on the SpeechDat II databases[1]. So far, experiments included the Slovenian, Spanish and German 1000 FDB SpeechDat (II) databases. Experiments have shown that the use of clustering algorithm results in a significant reduction of the number of triphones with minor degradation of word accuracy.

1. INTRODUCTION

The development of speech technology in the last few years raised an interest in the research of the multilingual speech recognition. In order to reduce the complexity of a multilingual recogniser and to reduce the cost of a cross-language transfer of speech technology, the development of methods for the definition of the multilingual phonetic inventories is of increasing concern.

The definition of the multilingual phonetic inventories by exploiting similarities among sounds of different languages is a promising approach. First attempt was reported in [1]. Here the multilingual phonetic inventory, consisting of language-dependent and language-independent speech units, was defined using the data-driven clustering technique. Other attempts based on different distance measures and clustering techniques also followed [2,3,4,5], however, all the work so far was focused on the context independent phoneme modelling (monophones). These experiments have shown that the transition from language dependent monophone set to multilingual inventory of monophones may result in a degradation of recognition accuracy due to the lack of acoustic resolution of the multilingual phoneme set.

The transition from the context independent to context dependent phoneme modelling seems inevitable in order to improve the performance of multilingual speech recognition systems, i.e. the speech recognisers with multilingual vocabulary. The development of a method for the definition of the multilingual set of context dependent phoneme models requires the definition of new clustering criteria.

In this paper, a clustering algorithm for the definition of multilingual set of context dependent phoneme models (triphones) is proposed. The clustering algorithm bases on a distance measure for triphones defined as the combination of explicit estimation of the similarity of the phonemes of left and right contexts and the central phonemes.

2. TRIPHONE DISTANCE MEASURE

The crucial problem concerning the use of triphone modelling is large number of triphone models, which requires large amounts of training data. Since the amount of training data is usually limited many of the triphone speech units are rarely or even never seen during the training. For this reason the direct implementation of the distance measures that were defined for the

[1] The use of SpeechDat databses was enabled by the Siemens AG and the Universitat Politecnica de Catalunya.

monophones, such as [1, 2, 3, 4] is not appropriate for the definition of multilingual set of triphones.

Our definition of the distance measure for triphones bases on the fact that the triphone is "a monophone in a certain context". Therefore, the similarity of two triphones can be estimated also indirectly - by explicitly estimating the similarity of both central phonemes, both left-context phonemes and both right-context phonemes. The similarity of two triphones l_1-c_1+r_1 and l_2-c_2+r_2 (l, c and r denote the left context - phoneme, right context - phoneme and the central phoneme, respectively) was therefore defined as:

$$S(l_1\text{-}c_1\text{+}r_1,\ l_2\text{-}c_2\text{+}r_2) = L\,s(l_1,l_2) + C\,s(c_1,c_2) + R\,s(r_1,r_2) \quad (1)$$

where s denotes the similarity of two phonemes, L, C, R are the weights for setting the influence of each phoneme - level similarity estimates, and $S(l_1\text{-}c_1\text{+}r_1,\ l_2\text{-}c_2\text{+}r_2)$ is the resulting similarity of both triphones.

Such definition of distance measure for triphones can be based on any type of phoneme-distance measure (s in Equation 1). In our case, the phone-distance measure was defined as suggested in [1]:

$$s(f_i,f_j)=s(f_j,f_i)= \frac{1}{2}\sum_{k=1}^{N}\left[c(f_i,f_k)+c(f_j,f_k)-\left|c(f_i,f_k)-c(f_j,f_k)\right|\right] \quad 1 \le i,j \le N,\ i \neq j \quad (2)$$

where $s(f_i,f_j)$ denotes the similarity between phonemes f_i and f_j, N is the number of phonemes, $c(f_i,f_k)$ is the number of confusions between phonemes i and phone j.

Described definition of distance measure for triphones has two major advantages. First it offers an accurate estimation of a triphone similarity (similarity of triphones is likely to be higher in a matching context and vice-versa). Next, such definition can provide a reliable estimation of similarity between triphones even in case of "rare" or "unseen" triphones.

3. CLUSTERING ALGORITHM

Having defined the distance measure for the triphones, the clustering algorithm for automatic identification of the triphones that are similar enough to be equated across the languages was defined.

A group of triphones is equated if an average distance among all triphones from the group is less than a predefined threshold T. Average distance among M triphones was defined as:

$$S(\varphi_1,\varphi_2,\mathrm{K},\varphi_M) = \frac{\sum_{k=1}^{M}\sum_{l=1}^{M} S(\varphi_k,\varphi_l)}{\sum_{k=1}^{M} k}$$

$$\varphi_k,\varphi_l \in (\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_M),\ k \neq 1 \quad (3)$$

where φ_k denotes the triphone l_k-c_k+r_k, (φ_1, φ_2, ... , φ_M is the group of triphones, S (φ_1, φ_2, ... , φ_M) is the average distance among all triphones from the group (φ_1, φ_2, ... , φ_M). To find all groups of triphones that complies with the condition from the Equation (3), the following 2-stage search algorithm was applied.

In the first stage, a list of most similar phonemes (poly-phonemes) was defined using the method described in [1]. A partial list of poly-phonemes covering all three languages is given in Table 1.

n	Slovene	German	Spanish
1	**a**	**a**	**a**
2	**O**	**O**	**o**
3	**n**	**n**	**n**
4	**l**	**l**	**l**
5	**t**	**t**	**t**
6	**m**	**m**	**m**

Table 1. A partial list of poly-phonemes for the Slovene, German and Spanish language.

In the second stage, the groups of triphones to be equated were identified. The search for these groups was limited to the classes of triphones consisting of triphones with the phonemes of the same poly-phoneme as the central phoneme. For example, the search for the similar triphones was first started among the triphones of all three languages with either Slovenian phoneme **a**, German phoneme **a** or Spanish phoneme **a** as the central phoneme. Next, the search for the groups of similar triphones continued among the triphones with either Slovenian phoneme **O**, German phoneme **O** or Spanish phoneme **o** as the central phoneme, etc. Such limitation of search has proven to significantly improve the convergence of the algorithm for the identification of the groups of similar triphones due to the large number of triphones

This clustering algorithm outputs the list of triphones that are similar enough to be equated across the languages. The unlisted triphones remain language specific. The degree of equated

triphones can be adjusted by the threshold *T*. The value of *T* was derived experimentally (values are given with the experimental results).

4. BASELINE RECOGNISER

The speech recognition system was based on HTK V2.1.1 with modified frontend module for enhancing the speech recognition robustness. The acoustic feature vector produced by the frontend module consisted of 24 mel-scaled cepstral, 12 Δ - cepstral, 12 ΔΔ - cepstral, high pass filtered energy, Δ - energy and ΔΔ - energy coefficients. This feature vector was processed using the algorithms for maximum likelihood channel adaptation [8] and linear discriminant analysis [8].

Such frontend module was chosen due to the results of previous tests on connected digits recognition task with 99 speakers of the Slovene speech database SNABI and tests on isolated digits recognition task with the databases SNABI and Voice-Mail (German).

The baseline speech recognition system consisted of three language specific recognisers (Slovene, German and Spanish) operating in parallel. The 3-state left-right topology was selected. The triphone models were initially built with 1 Gaussian mixture component per state. All together 24173 triphone models were defined (Sl.-7146, Ge.-12279,Sp.-4748). Parameter tying using the tree-based clustering algorithm (as implemented in the HTK) reduced the number of triphone models to 13074 (Sl.-3517, Ge.-6517,Sp.-3040). At the end the number of Gaussian mixture components per state was augmented to 32.

In the multilingual experiments, the three language specific recognisers operated in parallel using either three language specific model sets or one multilingual set of triphones where many of language specific triphones are tied and used by all three recognisers.

5. SPEECH DATABASES

The experiments were carried out using the speech databases produced in the framework of the SpeechDat II project [7]. These databases provide a realistic basis for developing voice driven teleservices and multilingual systems. The following SpeechDat databases were used:

- Slovenian 1000 FDB SpeechDat(II) [6],
- German 1000 FDB SpeechDat(II),
- Spanish 1000 FDB SpeechDat(II).

In all cases, the corpuses contained utterances of 1000 speakers. 800 speakers were used for the training and the remaining 200 speakers were used for the testing of the system. In all experiments the train and test sets were defined as recommended in SpeechDat II project specification.

Only 80 - 95 % of all utterances were useful for the experiments. Remaining utterances were skipped due to the following reasons:

- unusual pronunciation of digits,
- incomplete utterances (speech was cut off at the beginning or end of the utterance),
- unexpected utterances (background noise, comments, ...).

The system was trained using all corpuses of the train set, while for the testing the corpuses W1-W4 of all three databases, containing phonetically reach words, were used (total of 2252 utterances containing 1960 different words).

6. EXPERIMENTAL RESULTS

The baseline recogniser was tested in monolingual and in multilingual mode of operation, where the three language specific recognisers operated in parallel.
The recogniser performance for the monolingual tests is given in the Table 2.a. The word accuracy (*WA*) is listed for each language. The performance of the recogniser using the triphone models with 1 Gaussian mixture component per state (models: tri1) was low. Augmenting the number of Gaussian mixture components to 32 (models: tri32) significantly improved the word accuracy. The transition from the monolingusl to the multilingual mode of operation (Table 2.b) did not significantly degrade the recognizer performance. In most cases the recognizer correctly recognized the language. Errors in language identification usually ocured for the words that were already misrecognized in the monolingual tests. Therefore the errors in language identification did not cause additional errors in word recognition. The language identification rate (*LI*) was high for both types of triphone models and the word accuracy of multilingual tests approximately equals to the average word accuracy of the monolingual tests.

a)

models	WA		
	SL	ES	DE
tri1	67.51%	78.58%	76.77%
tri32	88.25%	93.91%	92.51%

b)

models	WA	LI
tri1	71.99%	91.61%
tri32	91.52%	93.10%

Table 2. The baseline recogniser performance for the monolingual tests (a) and for the multilingual tests (b).

Experiments with multilingual set of triphones were carried out for the recogniser with 13074 models and 1 Gaussian mixture component per state. The word accuracy was therefore much lower than it would have been in the case of models with 32 Gaussian mixture components per state. However, the purpose of the experiments was to determine the optimal values of the clustering parameters (weights *L,C,R* and threshold *T*) and to compare the performance of the multilingual triphone set to the performance of monolingual triphone sets running in parallel. Augmenting the number of Gaussian mixture components per state from 1 to 32 would improve the performance of the multilingual triphone set in the similar way as it did for the monolingual triphone set (Table 2).

The clustering algorithm was started at different values of weights *L,C* and *R* (see Equation 1) and at different threshold values (*T*) producing the multilingual triphone sets of different sizes. The performance of the recogniser using various multilingual triphone sets is given in the Tables 3.a, 3.b and 3.c.

Beside the word accuracy and the language identification rate, the global compression rate [4] was also followed. The global compression rate (*GCR*) was defined as:

$$GCR = \sum_{i=1}^{N} c_i \frac{M_i}{T_i} \qquad (4)$$

where L is the number of languages, T_i is the number of trainable models in language i, M_i is the number of merged models in language i and c_i is the ratio between the number of trainable models in language i and the number of trainable models in L languages.

The weights *L,C* and *R* were first set to the the values *L=1,C=0,R=1* (Table 3.a) . This way the similarity of both central phonemes did not have any influence to the resulting similarity of both triphones. The search for the groups of similar triphones was limited to the classes of triphones consisting of triphones with the phonemes of the same poly-phoneme as the central phoneme. Therefore the similarity of both central phonemes has already been considered during the search for the groups of similar triphones.

The use of multilingual set of triphones (models: tri1C) produced at weights *L=1,C=0,R=1* can reduce the total number of triphones of the baseline system (models: tri1), but it also results in a decrease of word accuracy and language identification rates in case of multilingual experiments (results from Tables 3 (WA-MULTI) are also shown on Figure 1). In best case the *GCR* of 24.19% is achieved at approximately 1% decrease of *WA* rate and more than 5% decrease of *LI* rate. Using the multilingual set of triphones for the monolingual experiments have shown an improvement of the word accuracy in case of Slovene language for the threshold values larger than 100 (results from Table 3.a (WA-SL) are shown also on Figure 1).

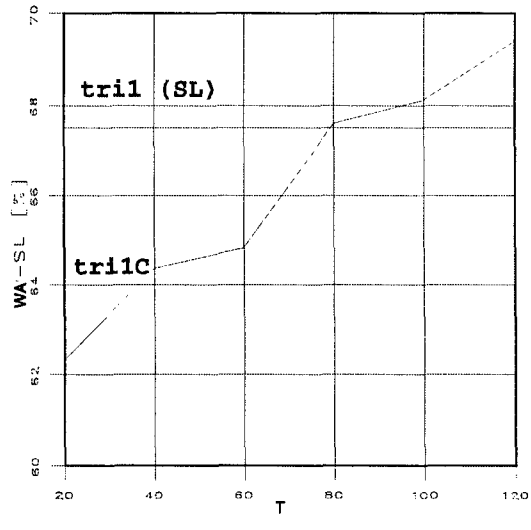

Figure 1. Word accuracy in case of monolingual experiments (Slovene language) using the multilingual set of triphones produced at various threshold values.

Next the value of weight *C* was increased to 0.5 (actual values of weights was *L,C, R was 2, 1, 2,* respectively, since only integer values were allowed). Increasing the value of weight *C* was

found to improve the performance of the recogniser with multilingual set of triphones

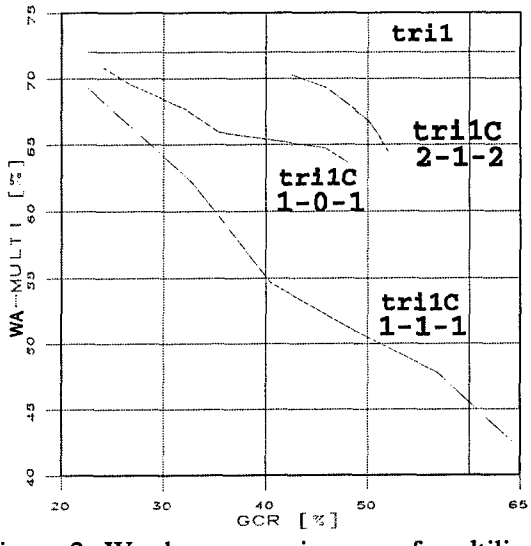

Figure 2. Word accuracy in case of multilingual experiments for various values of weights *L,C* and *R* as a function of *GCR*.

(Table 3.b). The WA and LI rates were similar as for the *C=0,* however the *GCR* was much higher (Figure 2). In this case the the *GCR* of 54.57% was achieved at approximately 1.6% decrease of *WA* rate and less than 5% decrease of *LI* rate. Such reduction of the total number of triphones with minor degradation of the *WA* can be considered as an improvement of the baseline system performance. As for the case of *C=0,* the use of multilingual set of triphones for the monolingual experiments improved the word accuracy in case of Slovene language for the threshold values of 400 or more.

Further increase of weight *C* (*C=1*) did not improve the performance of the recogniser with multilingual set of triphones (Table 3.c). Setting the *C* to 1 can produce the multilingual set of triphones with the highest *GCR,* but on the other hand, it significantly reduces the *WA* and *LI* (Figure 2).

a) *L=1,C=0,R=1*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	20	6498	62.34%	65.92%	69.51%	63.68%	75.73%	47.99%
tri1C	40	6799	64.38%	67.34%	70.73%	64.74%	77.57%	45.79%
tri1C	60	8226	64.83%	68.23%	72.23%	65.94%	78.23%	35.38%
tri1C	80	8662	67.60%	69.81%	73.90%	67.63%	80.37%	32.19%
tri1C	100	9424	68.12%	70.10%	74.95%	69.57%	84.52%	26.63%
tri1C	120	9758	69.41%	72.67%	75.85%	70.84%	86.07%	24.19%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

b) *L=2,C=1,R=2*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	100	5942	63.36%	63.21%	72.11%	64.43%	78.95%	52.04%
tri1C	160	6026	65.29%	64.02%	73.57%	65.14%	79.22%	51.43%
tri1C	180	6068	66.40%	64.89%	74.81%	65.63%	79.75%	51.12%
tri1C	260	6208	66.91%	66.12%	75.98%	66.82%	81.21%	50.10%
tri1C	340	6784	67.73%	69.47%	76.51%	69.27%	84.78%	45.89%
tri1C	400	7239	69.23%	73.20%	76.68%	70.32%	86.67%	42.57%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

c) *L=1,C=1,R=1*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	120	4238	29.12%	38.95%	40.72%	42.35%	58.89%	64.47%
tri1C	140	5284	28.63%	45.81%	47.34%	47.83%	63.55%	56.84%
tri1C	180	6475	37.78%	52.73%	53.59%	51.21%	69.26%	48.15%
tri1C	200	7526	46.62%	57.37%	59.45%	54.67%	76.31%	40.48%
tri1C	240	8577	56.16%	62.48%	64.83%	62.13%	82.74%	32.81%
tri1C	280	9971	65.41%	69.41%	71.73%	69.25%	86.07%	22.64%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

Table 3. Performance of the recogniser using various multilingual sets of triphones produced at different values of weights *L,C,R.*.

7. CONCLUSION AND FUTURE WORK

Experiments have shown that the use of clustering algorithm can produce the multilingual set of triphones that achieves almost the same word accuracy as the language specific triphone sets operating in parallel. Slight decrease of the word accuracy is acceptable considering the fact that the number of triphones in a multilingual set of triphones is significantly smaller than total number of triphones in the language specific triphone sets. In best case the use of clustering algorithm resulted in a reduction of the number of triphones by more than 40% with degradation of word accuracy by 1.67%: Such result shows that the multilingual set of triphones produced by the clustering algorithm can improve the performance of a multilingual recogniser based on language specific triphone sets operating in parallel.

The monolingual experiments with multilingual set of triphones have shown that in some cases the use of multilingual set of triphones can also improve the performance of the monolingual recognisers, that is, the performance of the recognisers based on monolingual triphone sets. Such improvement has been observed for the Slovenian language where the performance of the recogniser using the Slovenian triphone set was significantly lower than the performance of the recogniser (based on the Spanish and German triphone sets) for the Spanish and German languages. The multilingual set of triphones tends to equalise the performance of all monolingual triphone sets that were used for definition of the multilingual triphone set.

Results of the monolingual experiments using the multilingual triphone set indicates that the multilingual triphone set might also perform well for the new languages, that is the languages that were not included during the definition of the multilingual triphone set. However, no experiments have been done so far to prove this.

In future, the number of SpeechDat databases will be increased in order to expand the scale of experiments and to provide more reliable assessment of the clustering algorithm efficiency.

The clustering algorithm bases on a definition of distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. In this case the monophone distance estimation method was based on the algorithm of Houtgast. In future, other methods of monophone distance estimation will be also considered.

8. REFERENCES

[1] O. Andersen, P. Dalsgaard and W. Barry, *Data-Driven Identification of Poly- and Mono-phonemes for four European Languages.* 1993, Proc. EUROSPEECH '93, Berlin, pp. 759 - 762

[2] J. Koehler, *Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds.* 1996, Proc. ICSLP '96, Philadelphia, pp. 1780 - 1783

[3] K. M. Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters.* Oregon Graduate Institute of Science & Technology, Dissertation, 1996

[4] P. Bonaventura, F. Gallocchio, G. Micca, *Multilingual Speech Recognition for Flexible Vocabularies.* 1997, Proc. Eurospeech '97, Rhodos

[5] F.Weng and H. Bratt and L. Neumeyer and A. *Stolcke, A study of Multilingual Speech Recognition.* 1997, Proc. Eurospeech '97, Rhodos

[6] JanezKaiser, Zdravko Kačič, *Development of the Slovenian SpeechDat.* Speech Database Development for Central and Eastern European Languages, Granada,1998

[7] H. Hoege, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, *European speech Databases for Telephone Applications.* 1997, Proc. ICASSP '97, Muenchen, pp. 1771 - 1774

[8] A. Haunstein, E. Marschall, *Methods for Improved Speech Recognition over the Telephone Lines.* Proceed. IEEE IC ASSP, 1999